An Introduction to Equating and Scaling and the New ACH and MAAS Reportable Scales

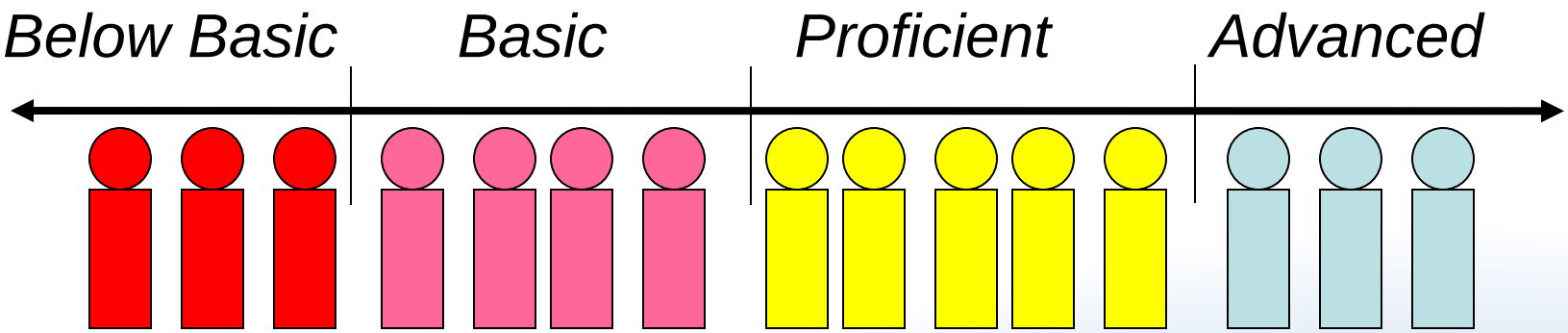
2010 LEAD Conference

Erika Hall, PhD

PEARSON

Context for Discussion

- Standard Setting was conducted in summer of 2010 for the new ACH and MAAS assessments
- Goal: Determine the cut scores or standards that define the different achievement levels



Context for Discussion

- Necessary to maintain the same standards for performance from one year to the next
  - Degree of ability required to be classified as "Proficient" or "Advanced" must be the equivalent from year to year.
- Provides for the common frame of reference to support standards-based inferences
  - AYP calculations
  - School/district/state-level improvement

How do we do this?

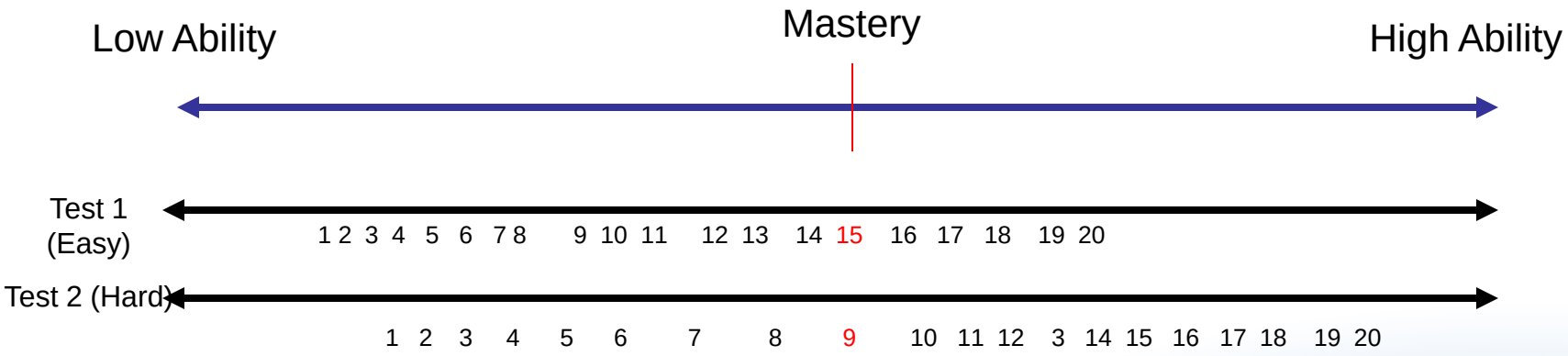
- Common Misconceptions
  - Cut scores are always defined with respect to a given number or percent correct.
  - Changes in the number of points required to achieve a given achievement level from one year to the next reflects a different standard for performance.
    - Raw score required to achieve proficiency on the 2010 Grade 3 Math test is 37, raw score required to achieve proficiency in 2011 test is 34.

Extreme Example

- Two 20 item mathematics tests intended to assess addition and subtraction.
  - Test 1: 20 easiest items in the bank
  - Test 2: 20 most difficult addition items in the bank
- Is it reasonable to use the same raw score to define proficiency on these two tests?
- Does a raw score of 15 reflect the same level of ability on Test 1 as it does on Test 2?
- Is a student who receives a 12 on Test 2 lower in ability than a student who receives a 16 on Test 1?

Context for Discussion (cont.)

- Standards are defined relative to a specific scale and delineate a degree of proficiency.



Context for Discussion (cont.)

- Need a process that corrects for differences in overall test difficulty so that student- and group-level results can be compared across forms on a common scale.
- Need a means of translating students results to a useful metric that facilitates the appropriate interpretation and use of student results

Equating and Scaling !

What is Equating?

- Equating is the statistical process of adjusting for differences in test form difficulty so scores resulting from different forms can be used interchangeably (Kolen & Brennan, 2004)
- Equating is necessary when the results of two assessments, built to the same content and statistical specifications but using different sets of items, are to be compared.

How is Equating Conducted?

Depends on a variety of factors

1. Data Collection Design
   - Depends on statistical and practical issues
2. Assumptions you are willing to make about your data and the equating relationship
3. Whether Item Response Theory (IRT) methods are applied

Data Collection Design

- Randomly Equivalent Groups Design
- Common-Item Nonequivalent Groups Design (Common Items)
- Single Group – with counterbalancing

Randomly Equivalent Groups Design

- Students randomly assigned to a test using a spiraling process.

  Assumptions:

  - Groups are comparable and equivalent due to spiraling
  - Groups are representative of the target test taking population
  - Differences in test performance are due solely to differences in test difficulty

  Constraint:

  - Build and produce all forms that need to be equated at the same time.
  - Need to expose to expose multiple operational forms at the same time.

Common Items Non-Equivalent Groups

- Test to be equated administered at two different times to two different populations
  - Cannot assume the two populations to be equivalent in ability
  - Differential performance on the test cannot be said to be solely due to difference in item difficulty.
- Develop tests to have a common item link
- Use these common items as the basis for determining the equating relationship.

  Assumptions
  - Difficulty of selected common items does not change from one administration to the next
  - Differential performance between two groups on common items provides estimate of difference in ability of groups.

Common Persons Design

- Same group of students takes both forms
- Must use counterbalancing so there is not a fatigue effect.

Problem:

  - Requires multiple testing for same set of kids
  - Exposes two forms worth of items

Basic Example – Randomly Equivalent Groups

	Average Score on Test	
	Group A (10,000)	**Group B (10,000)**
Form X (50 items)	40	
Form Y (50 items)		44

- Form Y is 4 items easier, on average, than Form X.
- Can equate form Y to form X, by removing the 4 point advantage associated with taking the easier form.

Basic Example REG – Mean Equating

- *m*X(Y)=Y-4

Score on Form X	Score on Form Y	Form Y Score after Equating to Form X
37	37	**33**
38	38	**34**
39	39	**35**
40	40	**36**
41	41	**37**
42	42	**38**
43	43	**39**
44	44	**40**
45	45	**41**
46	46	**42**

Linear Equating

Argue that the effect of difference in test difficulty differs along the score scale.

- Scores that are equal distance from their respective means are considered equivalent.
- Form X = Mean of 40, SD=5
- Form Y = Mean of 44, SD=3

  Score of 41 on Y is one SD below mean

  Score of 35 on X is one SD below mean

  So Lx(41)=35

Linear Equating Conversion Table

Score on Form X	# of Deviations from Mean	Score on Form Y	# of Deviations from Mean	Ly(x)
34	-1.20	34	-3.3	23.33
35	-1.00	35	-3.0	25.00
36	-0.80	36	-2.7	26.67
37	-0.60	37	-2.3	28.33
38	-0.40	38	-2.0	30.00
39	-0.20	39	-1.7	31.67
40	0.00	40	-1.3	33.33
41	0.20	41	-1.0	35.00
42	0.40	42	-0.7	36.67
43	0.60	43	-0.3	38.33
44	0.80	44	0.0	40.00
45	1.00	45	0.3	41.67
46	1.20	46	0.7	43.33
47	1.40	47	1.00	45.00

Scaling

- In this case the common metric that facilitates comparisons is the *Form X raw score metric*.
- Using such a metric is confusing, however, since equated raw scores are often non-integer values. Hard to report a 31.67!
- For this reason we typically translate the base form raw score metric (Form X)to a score scale that is easier to report and interpret.
- The score scale is maintained across forms through the equating process

Scale Scores

- Definition: Transformation of an estimate of ability (e.g., raw score/theta) to a metric that enhances the interpretability of scores.
- How established?
  - Using normative information
  - In light of a defined standard
- How operationalized?
  - Typically a linear transformation equation of the form

    A (Estimate of Ability) + B= Scaled Score

Back to Earlier Example

Normative Scaling

- Test X has a mean of 40 and SD=5
- Establish scaling equation so that mean performance on Form X = 300, SD=50

Scaled Score = $\left(\frac{x-40}{5}\right)*50+300$

Apply Scaling Equation to Equated Results

Form X	Form X Scaled Score	Form Y	L(y)x	Form Y Scaled Score
34	240	34	23.33	133
35	250	35	25.00	150
36	260	36	26.67	167
37	270	37	28.33	183
38	280	38	30.00	200
39	290	39	31.67	217
40	300	40	33.33	233
41	310	41	35.00	250
42	320	42	36.67	267
43	330	43	38.33	283
44	340	44	40.00	300
45	350	45	41.67	317
46	360	46	43.33	333
47	370	47	45.00	350

Non –Equivalent Groups Example

- More complicated
  - Test form X administered in 2010
  - Test form Y administered in 2011
- Scale established in 2010 using form X
- 30% of items on Form Y are common to Form X (a.k.a. anchor set, linking set)
- Use of Item Response Theory

Item Response Theory

- Mathematical model that provides for:
  - estimates of item difficulty and estimates of student ability that are on a *common scale – theta scale*
  - *a* means for estimating the probability that a student at a given ability level will answer a given item correctly.
- Can aggregate these probabilities over items to establish a predicted raw score on a given test.

Test Characteristic Curve

Figure 1. MAAS Grade 3 Math TCC

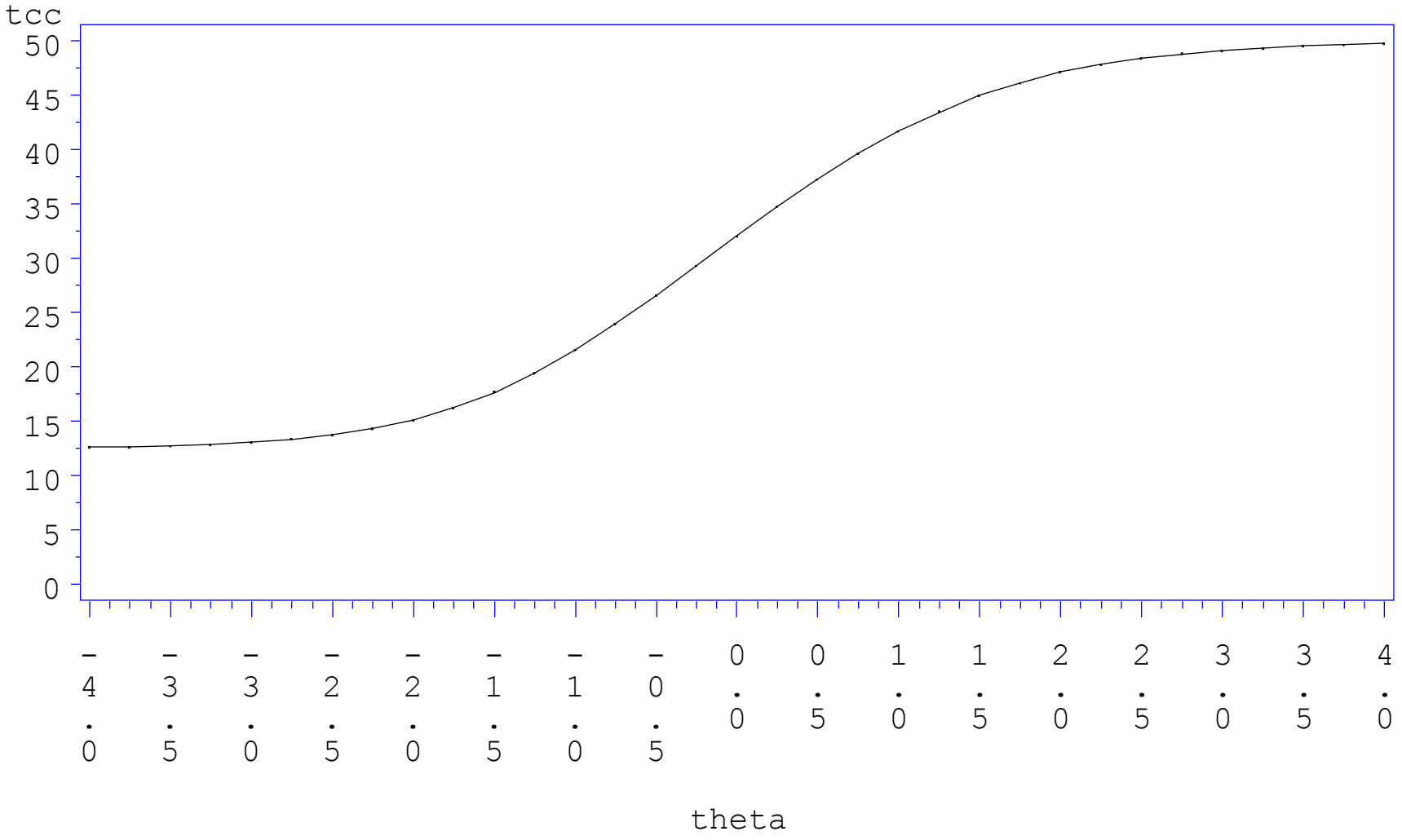
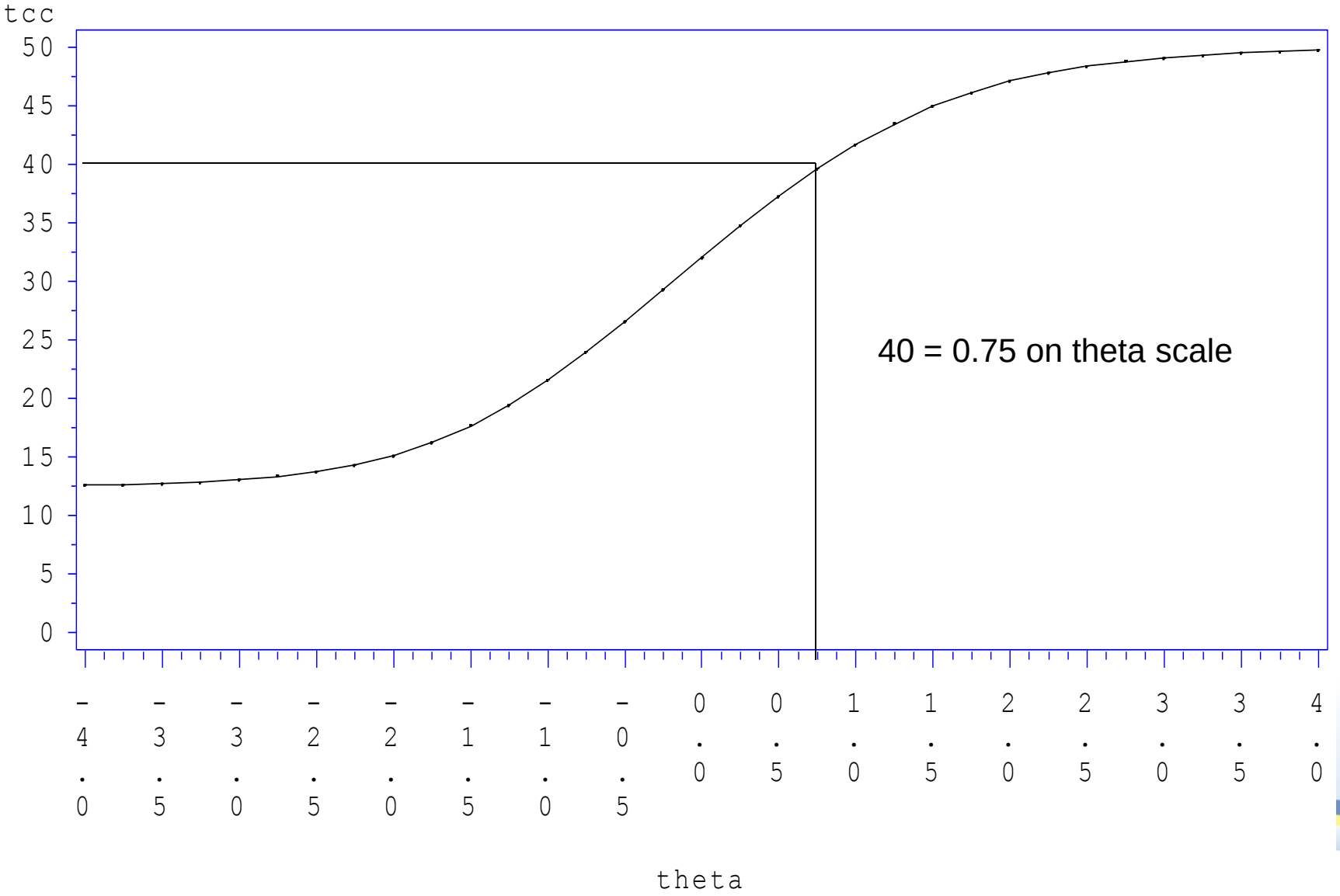


Figure 1. MAAS Grade 3 Math TCC



Form X Raw Score to Scaled Score Table

Form X		
Raw Score	Ability Estimate	Scaled Score
29	-0.25	288
30	-0.15	293
31	0.00	300
32	0.10	305
33	0.22	311
34	0.35	318
35	0.50	325
36	0.64	332
37	0.68	334
38	0.71	336
39	0.73	337
40	0.75	338
41	0.85	343
42	1.00	350
43	1.11	356
44	1.25	363

CINEG Equating

- Common item parameters should not change from one use to the next – parameters are sample dependent.
- The extent that they do reflects a shift in the overall ability of the groups assessed
- Use common items to determine the "equating constants" necessary to place the item parameters for the new form (Y) on the scale of the old form (X).
- Once have equated Form Y parameters can estimate the ability associated with each raw score (as previously discussed)

Raw Score to Scale Score Conversion Table

	Form X		Form Y	
Raw Score	Ability Estimate	Scaled Score	Ability Estimate	Scaled Score
29	-0.25	288	-0.23	289
30	-0.15	293	-0.15	293
31	0.00	300	-0.08	296
32	0.10	305	0.00	300
33	0.22	311	0.24	312
34	0.35	318	0.50	325
35	0.50	325	0.58	329
36	0.64	332	0.75	338
37	0.68	334	0.78	339
38	0.71	336	0.84	342
39	0.73	337	0.90	345
40	0.75	338	1.10	355
41	0.85	343	1.14	357
42	1.00	350	1.20	360
43	1.11	356	1.24	362
44	1.25	363	1.35	368

ACH Scaled Score Metric

- Range 600-900
- Mean of 2010 test taking population set to 750, SD=35 for each test

Individual Score	Normative Score Interpretation (Regardless of Administration)
750	Performance equivalent to average performance of 2010 test taking population
715	Performance equivalent to one standard deviation below mean of 2010 test taking population

Things to keep in mind

- To make criterion referenced interpretations you need to know the cut scores associated with each achievement level for each test.
  - Compare a students score to a given cut score to see how far they are away from achieving proficiency.
- New ACH scale is completely different from previous ACH scale so comparisons to previous years ACH scaled scores are not appropriate.
- Mathematics and Reading Language Arts tests are no longer on a vertical scale

MAAS Scaled Score Metric

- Range 200-500
- Mean of 2010 test taking population set to 300, SD=50 for each test

Individual Score	Normative Score Interpretation (Regardless of Administration)
300	Performance equivalent to average performance of 2010 test taking population
250	Performance equivalent to one standard deviation below mean of 2010 test taking population

Questions?

Erika.Hall@Pearson.com

www.curesearch.org